AIRPLANE PRICE PREDICTION

Abstract

Airplane pricing plays a crucial role in the aviation industry, influencing decisions for airlines, leasing companies, manufacturers, and buyers. Estimating an airplane's price is complex due to various factors such as manufacturer, model, engine type, seating capacity, aircraft age, fuel efficiency, and maintenance history. This project will develop a machine learning model to predict airplane prices based on these key attributes.

We will begin with data preprocessing, ensuring accuracy by handling missing values and inconsistencies. Exploratory Data Analysis (EDA) will help identify significant price-influencing factors, such as engine type, age, capacity, and fuel consumption. A Multiple Linear Regression model will be trained on historical data, with performance evaluated using R-squared (R^2), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). If needed, alternative models will be explored to improve accuracy

By combining data analysis with machine learning, this project aims to create a useful tool for predicting airplane prices, helping various stakeholders make informed and data-driven decisions in the aviation industry

TABLE OF CONTENTS

CHAPTER - I

INTRODUCTION

1.1 Scope of Analysis

The scope of this analysis is to develop a predictive model for estimating airplane prices using Multiple Linear Regression. The dataset consists of 12,377 airplane records with various attributes, such as Num of Engines, Capacity, Range in km, Fuel Consumption, Maintenance Costs, and categorical features like Engine Type, Sales Region, and Model of the airplane.

Exploratory Data Analysis (EDA) will be conducted to uncover patterns and relationships within the dataset, utilizing various charts and visualizations to aid interpretation. The analysis involves preprocessing the data by handling categorical variables. A dashboard will be created to present key insights interactively. The regression model is built using the lm() function in R.

The expected outcome is to identify key factors influencing aircraft prices, measure prediction accuracy, and provide insights for aircraft buyers and sellers through visualizations and an interactive dashboard.

1.2 Approach of Analysis

The analysis for predicting airplane prices will follow a structured approach, starting with data cleaning and preprocessing. This includes handling missing values, encoding categorical variables, and scaling numerical features to ensure consistency.

Exploratory Data Analysis (EDA) will be conducted to identify patterns and relationships between key factors like engine type, capacity, and range. Various visualizations, including bar charts and line charts, will help interpret the data. Interactive dashboards will be created to showcase trends and comparisons.

Next, the dataset will be split into training and testing sets, and a supervised learning algorithm, specifically Multiple Linear Regression, will be used to build the model. Performance will be evaluated using metrics like R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Accuracy Percentage.

This structured approach will help identify key factors influencing airplane prices, improve prediction accuracy, and provide valuable insights for aircraft buyers and sellers through interactive dashboards and visual analysis.

CHAPTER - II

GATHERING DATA

2.1 Dataset Description

Load necessary Packages

tidyverse – A collection of R packages for data manipulation, visualization, and modeling, including ggplot2, dplyr, tidyr, readr, and more.

caret – A package used for machine learning, model training, and evaluation, particularly useful for regression and classification tasks.

```{r}
library(tidyverse)
library(caret)
```

Load the Dataset

The **read.csv()** function in R is used to read CSV (Comma-Separated Values) files into a data frame.

```{r}
dataset=read.csv("C:/Users/deepi/Desktop/deeps/airplane_price_dataset.csv")
```

Dimension of data

Dim() function returns the dimensions of the dataset which is count of Rows and Columns. It is commonly used to check the size of a data frame or matrix.

```{r}
dim(dataset)
```

```
[1] 12377    11
```

Airplane Dataset

This Airplane Price dataset contains 12,377 rows and 11 columns, initially written in Turkish and Later translated into English.

	Model	Üretim.Yılı	Motor.Sayısı	Motor.Türü	Kapasite	Menzil..km.	Yakıt.Tüketimi..L.saat.	Saatlik.Bakım.Maliyeti....	Yaş	Satış.Bölgesi	Fiyat....
1	Bombardier CRJ200	1987	2	Turbofan	50	3000	14.36	2185.43	36	Asya	12857083.9
2	Bombardier CRJ200	1997	2	Turbofan	50	3000	4.03	1202.08	26	Avrupa	13914058.6
3	Airbus A320	1988	2	Turbofan	180	6300	13.26	761.38	35	Avustralya	90735695.9
4	Boeing 737	2023	2	Turbofan	162	5700	14.61	592.63	0	Avustralya	136659689.4
5	Cessna 172	1985	1	Piston	4	1285	18.49	4245.99	38	Güney Amerika	203798.1
6	Airbus A350	1982	2	Turbofan	350	14800	8.82	1869.09	41	Asya	354976612.8
7	Boeing 737	1993	2	Turbofan	162	5700	5.79	2443.75	30	Avrupa	57032326.6
8	Cessna 172	2018	1	Piston	4	1285	24.65	1814.65	5	Asya	396750.4
9	Cessna 172	1992	1	Piston	4	1285	41.41	2552.31	31	Avustralya	215144.5
10	Bombardier CRJ200	2014	2	Turbofan	50	3000	2.60	3165.87	9	Afrika	16981571.2
11	Airbus A320	2008	2	Turbofan	180	6300	4.22	3233.95	15	Güney Amerika	103115871.8
12	Boeing 777	1980	2	Turbofan	396	15600	7.72	3579.05	43	Avrupa	385983616.6
13	Bombardier CRJ200	2007	2	Turbofan	50	3000	8.44	1049.17	16	Kuzey Amerika	17772783.2
14	Boeing 777	1989	2	Turbofan	396	15600	8.76	1902.70	34	Avrupa	349042400.5
15	Boeing 777	1986	2	Turbofan	396	15600	4.40	2960.20	37	Asya	426940813.4
16	Airbus A350	1986	2	Turbofan	350	14800	4.55	898.22	37	Kuzey Amerika	378696823.9
17	Boeing 777	2018	2	Turbofan	396	15600	6.23	703.52	5	Kuzey Amerika	609835688.0
18	Boeing 737	2009	2	Turbofan	162	5700	6.64	4229.32	14	Güney Amerika	84803923.4
19	Boeing 737	2004	2	Turbofan	162	5700	9.06	1764.21	19	Asya	81240277.7
20	Cessna 172	1998	1	Piston	4	1285	49.48	835.48	25	Avustralya	205472.1
21	Cessna 172	2003	1	Piston	4	1285	17.95	3975.10	20	Güney Amerika	282880.8
22	Airbus A320	1984	2	Turbofan	180	6300	2.96	3970.72	39	Asya	85658440.2
23	Bombardier CRJ200	1994	2	Turbofan	50	3000	3.51	2113.10	29	Kuzey Amerika	12475000.6
24	Boeing 737	1994	2	Turbofan	162	5700	2.83	1989.04	29	Asya	69650507.5
25	Airbus A350	1997	2	Turbofan	350	14800	6.23	1899.42	26	Afrika	402568177.0
26	Bombardier CRJ200	2003	2	Turbofan	50	3000	8.14	4492.46	20	Avrupa	16455191.8

Variable Description

1. Model:

   Airplane model, influencing price based on brand and performance capabilities.

2. Üretim Yılı (Production Year):

   The year the airplane was manufactured (1980 – 2023).

3. Motor Sayısı (Number of Engines):

   Total number of engines on the airplane (1 for piston engines, 2 for others).

4. Motor Türü (Engine Type):

   Type of engine, such as Turbofan or Piston. Different engine types affect performance, fuel efficiency, and price.

5. Kapasite (Capacity):

   Passenger capacity of the airplane. Larger airplanes with higher capacity tend to have higher base prices.

6. Menzil (km) (Range in km):

   Maximum range the airplane can travel without refueling. Greater range increases operational flexibility and cost.

7. Yakıt Tüketimi (L/saat) (Fuel Consumption in L/hour):

   Average fuel consumption of the airplane. Turbofan engines consume less than piston engines, influencing operating costs.

8. Saatlik Bakım Maliyeti ($) (Hourly Maintenance Cost):

   The average hourly maintenance cost in USD. Higher maintenance costs can lower demand and impact pricing.

9. Yaş (Age):

   The age of the airplane, calculated as 2023 - Production Year. Older airplanes tend to have lower prices due to depreciation.

10. Satış Bölgesi (Sales Region):

    The region where the airplane is being sold (e.g., Asia, Europe, North America). Regional demand and currency variations can affect prices.

11. Fiyat ($) (Price in USD):

    The final price of the airplane, serving as the target variable for prediction.

2.2 Understanding Data

Summary of the Data

The Summary function provides insights into variable types, min/max values, mean, median, and percentiles. It identifies Categorical and Continuous variables.

```{r}
summary(data)
```

```
    Model            Üretim.Yılı    Motor.Sayısı    Motor.Türü           Kapasite      Menzil..km.    Yakıt.Tüketimi..L.saat.
 Length:12377       Min.   :1980   Min.   :1.000   Length:12377       Min.   :  4.0   Min.   : 1285   Min.   : 2.00
 Class :character   1st Qu.:1990   1st Qu.:2.000   Class :character   1st Qu.: 50.0   1st Qu.: 3000   1st Qu.: 5.95
 Mode  :character   Median :2001   Median :2.000   Mode  :character   Median :162.0   Median : 5700   Median : 9.82
                    Mean   :2001   Mean   :1.835                      Mean   :190.4   Mean   : 7782   Mean   :12.08
                    3rd Qu.:2013   3rd Qu.:2.000                      3rd Qu.:350.0   3rd Qu.:14800   3rd Qu.:13.47
                    Max.   :2023   Max.   :2.000                      Max.   :396.0   Max.   :15600   Max.   :49.97
 Saatlik.Bakım.Maliyeti....      Yaş        Satış.Bölgesi        Fiyat....
 Min.   : 500               Min.   : 0.00   Length:12377       Min.   :   145815
 1st Qu.:1627               1st Qu.:10.00   Class :character   1st Qu.: 14096814
 Median :2744               Median :22.00   Mode  :character   Median : 83921914
 Mean   :2744               Mean   :21.52                      Mean   :198833650
 3rd Qu.:3849               3rd Qu.:33.00                      3rd Qu.:384323881
 Max.   :5000               Max.   :43.00                      Max.   :978213229
```

There are totally 11 variables in the Airplane price dataset.

Categorical Variables:

- Model (char)
- Engine type (char)
- Sales Region (char)

Continuous variables:

- Year (int)
- Number of Engines (int)
- Capacity (int)
- Range (int)
- Fuel consumption (int)
- Maintenance cost (int)
- Price (int)

Structure of the Data

Str() function shows the structure of the data along with the datatypes, and sample of the data. It shows that the dataset contains two types of data which is Numeric and Character variables.

```{r}
str(dataset)
```

```
'data.frame':   12377 obs. of  11 variables:
 $ Model                    : chr  "Bombardier CRJ200" "Bombardier CRJ200" "Airbus A320"
"Boeing 737" ...
 $ Üretim.Yılı              : int  1987 1997 1988 2023 1985 1982 1993 2018 1992 2014 ...
 $ Motor.Sayısı             : int  2 2 2 2 1 2 2 1 1 2 ...
 $ Motor.Türü               : chr  "Turbofan" "Turbofan" "Turbofan" "Turbofan" ...
 $ Kapasite                 : int  50 50 180 162 4 350 162 4 4 50 ...
 $ Menzil..km.              : int  3000 3000 6300 5700 1285 14800 5700 1285 1285 3000 ...
 $ Yakıt.Tüketimi..L.saat.  : num  14.36 4.03 13.26 14.61 18.49 ...
 $ Saatlik.Bakım.Maliyeti....: num  2185 1202 761 593 4246 ...
 $ Yaş                      : int  36 26 35 0 38 41 30 5 31 9 ...
 $ Satış.Bölgesi            : chr  "Asya" "Avrupa" "Avustralya" "Avustralya" ...
 $ Fiyat....                : num  1.29e+07 1.39e+07 9.07e+07 1.37e+08 2.04e+05 ...
```

First few rows

The head() function in R is used to display the first few rows of a dataset. By default, it shows the first 6 rows.

```{r}
head(dataset)
```

Description: df [6 × 11]

	Model <chr>	Üretim.Yılı <int>	Motor.Sayısı <int>	Motor.Türü <chr>	Kapasite <int>
1	Bombardier CRJ200	1987	2	Turbofan	50
2	Bombardier CRJ200	1997	2	Turbofan	50
3	Airbus A320	1988	2	Turbofan	180
4	Boeing 737	2023	2	Turbofan	162
5	Cessna 172	1985	1	Piston	4
6	Airbus A350	1982	2	Turbofan	350

6 rows | 1-6 of 11 columns

CHAPTER - III

PREPARING & EXPLORING DATA

3.1 Data Exploration

Data exploration is the process of analyzing a dataset to understand its structure, patterns, and potential issues before applying machine learning models. It helps in making informed decisions about data cleaning, visualize relationships of the variables, model selection.

Data Visualization

Data visualization is the process of representing data graphically to help identify patterns, trends, and insights. It uses visual elements like charts, graphs, and maps to make complex data more understandable and accessible.

Charts

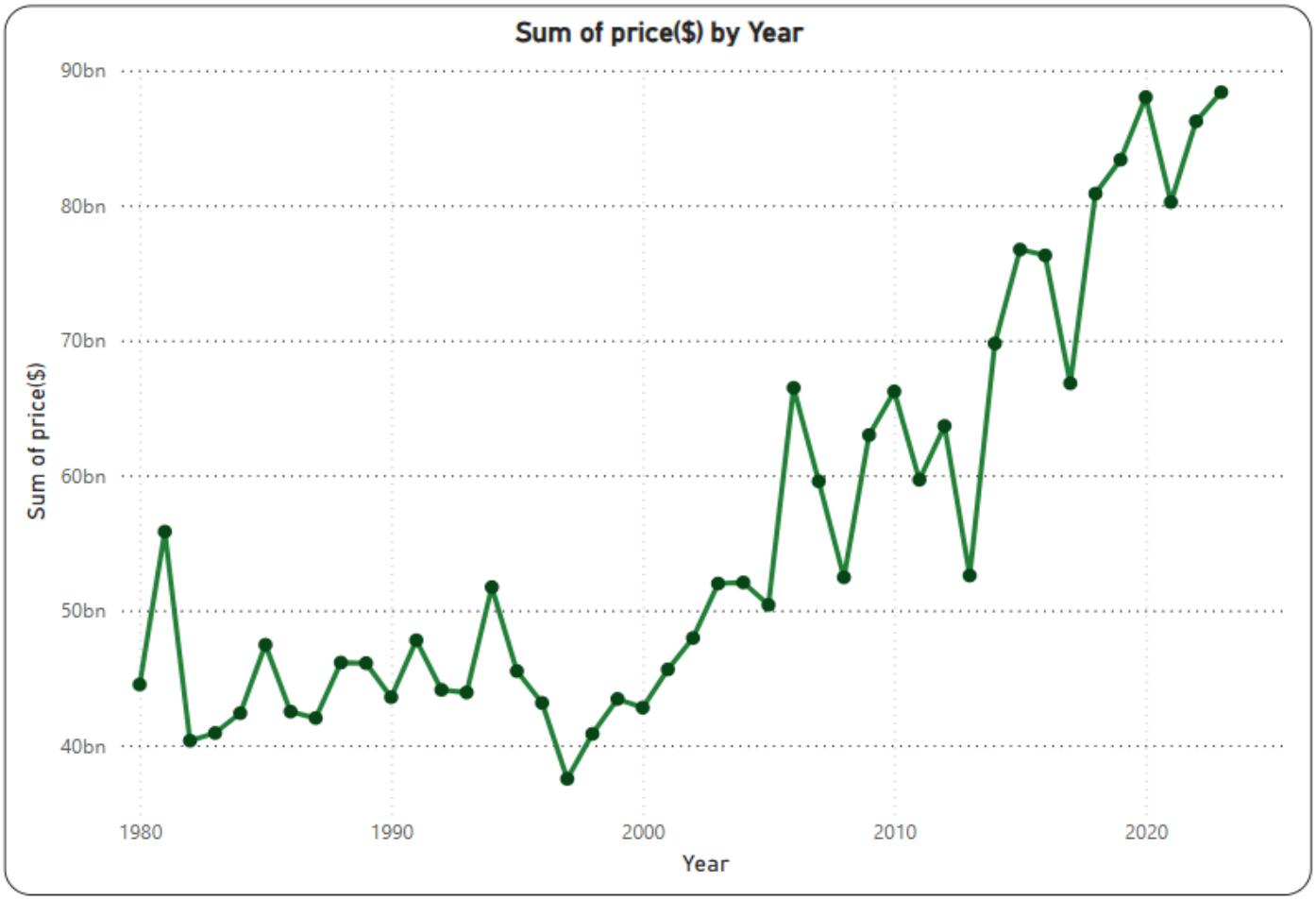
- **Line Chart:** Total Airplane Prices Over the Years

- **Pie Chart:** Fuel Consumption by Aircraft Model per Hour

- **Bar Chart:** Minimum Passenger Capacity by Aircraft Model

Line Chart:

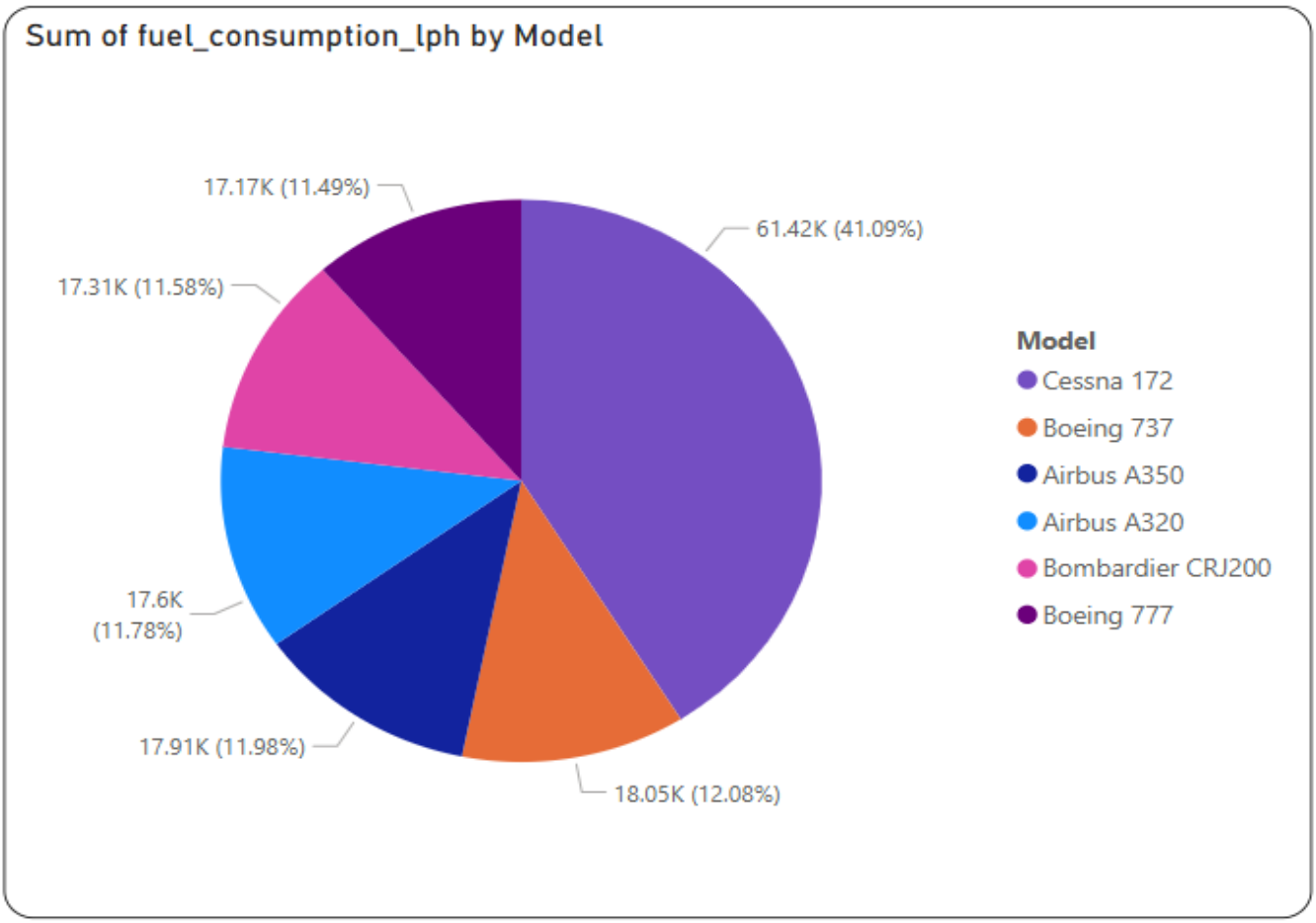
Total Airplane Prices Over the Years



The chart shows that airplane prices have gone up over time. There are some ups and downs, but overall, prices have increased a lot, especially after 2000. In recent years, prices are the highest they have ever been. This could be because more planes are being sold or they are getting more expensive.

Pie Chart:

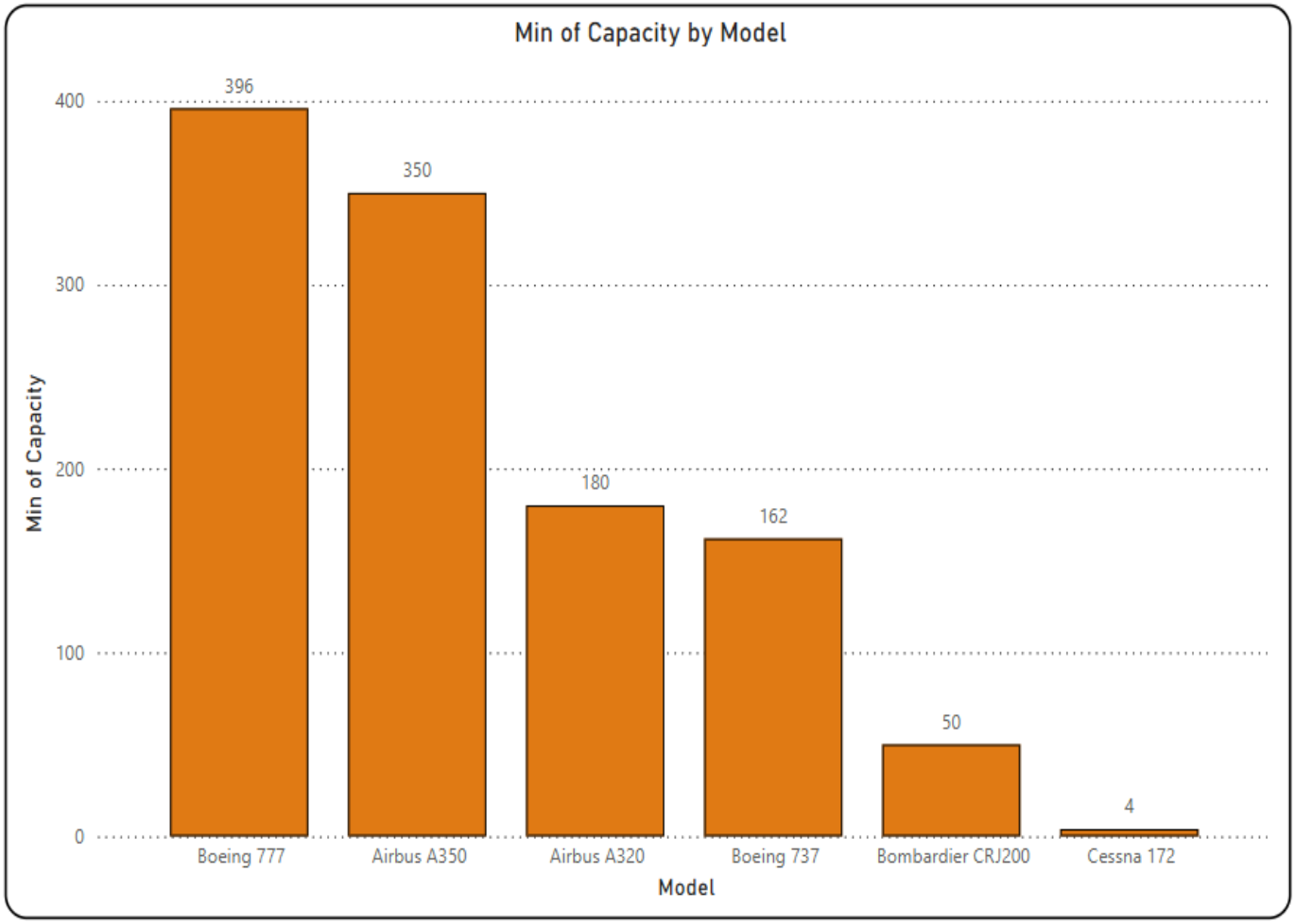
Fuel Consumption by Aircraft Model per Hour



This Pie Chart shows total fuel usage (liters per hour) across different aircraft. The Cessna 172 dominates with 61.42K LPH (41.09%), followed by the Boeing 737 at 18.05K LPH (12.08%). The Airbus A350, Airbus A320, Bombardier CRJ200, and Boeing 777 each consume around 17K-18K LPH (11.49%-11.98%). This highlights the Cessna 172 as the most fuel-intensive model.

Bar Chart:

Minimum Passenger Capacity by Aircraft Model



This bar chart compares the minimum passenger capacity of various aircraft. The Boeing 777 (396) and Airbus A350 (350) have the highest capacities, while the Airbus A320 (180) and Boeing 737 (162) fall in the mid-range. Smaller aircraft like the Bombardier CRJ200 (50) and Cessna 172 (4) have significantly lower capacities, highlighting the variation in seating among different models.

3.2 Issues in the dataset

Checking missing values

Colsums() function is used to count the number of missing values per columns in the data. As per the result the dataset does not contain missing values or null values.

```{r}
colSums(is.na(data))
```

```
                 Model                   Year             Num_Engines             Engine_Type                Capacity 
                     0                      0                       0                       0                       0 
              Range_km Fuel_Consumption_Lph Maintenance_Cost_per_Hour                     Age            Sales_Region 
                     0                      0                       0                       0                       0 
                 Price 
                     0 
```

Columns Names

Display the column names with the help of colnames() function.

```{r}
colnames(data)
```

```
 [1] "Model"                "Üretim.Yılı"          "Motor.Sayısı"         "Motor.Türü"          
 [5] "Kapasite"             "Menzil..km."          "Yakıt.Tüketimi..L.saat." "Saatlik.Bakım.Maliyeti...."
 [9] "Yaş"                  "Satış.Bölgesi"        "Fiyat...."           
```

- Column names were originally in Turkish, so It should be renamed in English for easier handling.
- Translated the column names to English.
- Use rename() function to change the column names to English.

3.3 Resolve Issues

Null values

The output from sum(is.na(data)) function confirms that there are no missing values in this dataset.

```{r}
sum(is.na(data))
```

```
[1] 0
```

Rename columns

we rename the variable names of our dataset using rename() from dplyr package , which were originally in Turkish, to English for easier interpretation and processing in our analysis.

```{r}
data <- dataset %>%rename(Model = Model,
Year =Üretim.Yılı ,
Num_Engines = Motor.Sayısı,
Engine_Type = Motor.Türü ,
Capacity = Kapasite,
Range_km = Menzil..km.,
Fuel_Consumption_Lph = Yakıt.Tüketimi..L.saat. ,
Maintenance_Cost_per_Hour = Saatlik.Bakım.Maliyeti....,
Age = Yaş,
Sales_Region = Satış.Bölgesi,
Price = Fiyat....)
```

```r
print("Column names in the Dataset")
colnames(dataset)
print("Column names after Alteration")
colnames(data)
```

```
 [1] "Column names in the Dataset"
  [1] "Model"
  [2] "Üretim.Yılı"
  [3] "Motor.Sayısı"
  [4] "Motor.Türü"
  [5] "Kapasite"
  [6] "Menzil..km."
  [7] "Yakıt.Tüketimi..L.saat."
  [8] "Saatlik.Bakım.Maliyeti...."
  [9] "Yaş"
 [10] "Satış.Bölgesi"
 [11] "Fiyat...."
 [1] "Column names after Alteration"
  [1] "Model"
  [2] "Year"
  [3] "Num_Engines"
  [4] "Engine_Type"
  [5] "Capacity"
  [6] "Range"
  [7] "Fuel_Consumption_Lph"
  [8] "Maintenance_Cost_per_Hour"
  [9] "Age"
 [10] "Sales_Region"
 [11] "Price"
```

The above code displays the original column names, along with the updated column names. Print() function used to print the command and colnames() function is used to print the names of the columns of both dataset which contains original column names and data which is updated to English.

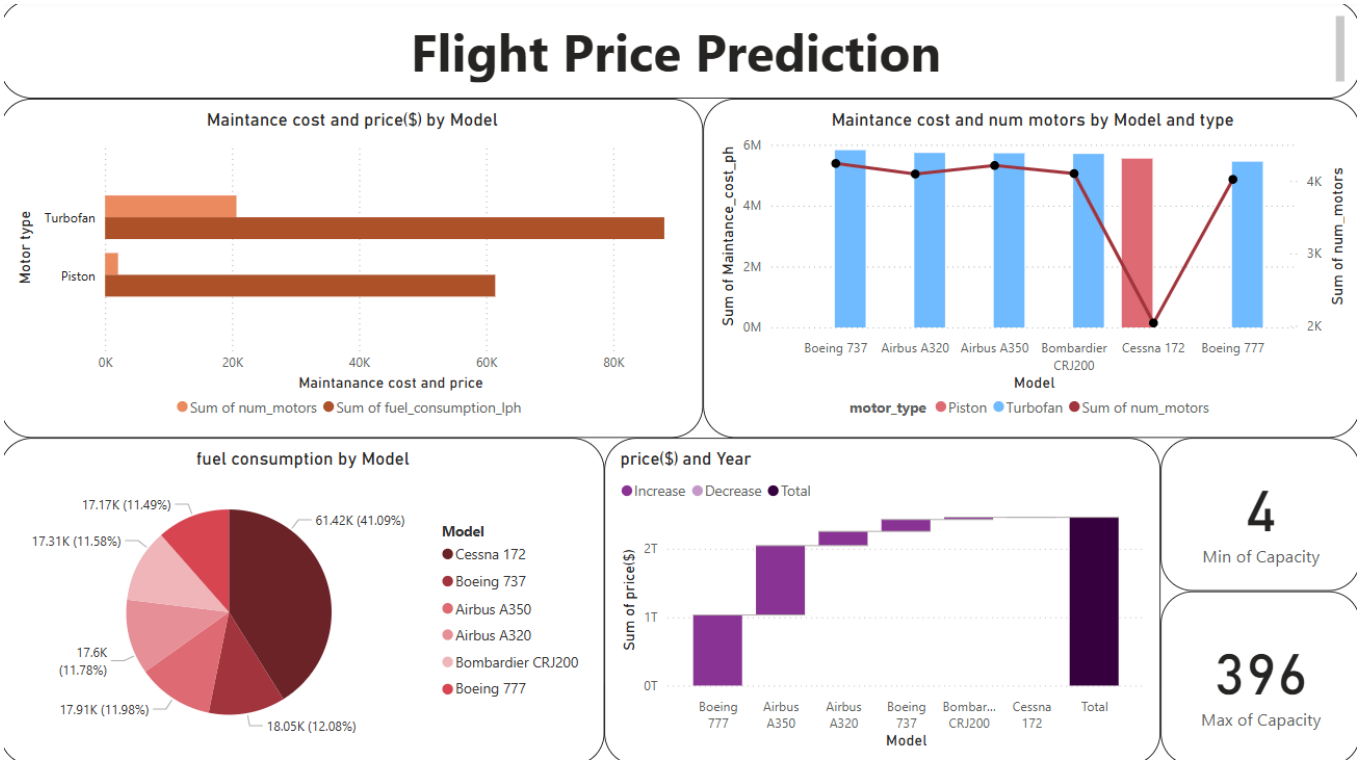
CHAPTER - IV

BUSINESS INTELLIGENCE INTERACTIVE DASHBOARDS

4.1 Dashboard Interpretation

A Dashboard is a visual interface that displays key information, metrics, and insights in a structured format. It helps users quickly analyze and monitor data using charts, tables, and interactive components.

Flight Price Analysis & Maintenance Insights Dashboard



This Power BI dashboard provides insights into flight prices, maintenance costs, fuel consumption, and engine types. It helps in understanding key cost drivers for different aircraft models.

Dashboard Insights:

1. Clustered Bar Chart: Maintenance Cost and Price by Model

- Compares maintenance cost and price based on motor type (Turbofan vs. Piston).
- Turbofan engines have higher fuel consumption and maintenance costs.

2. Line & Clustered Column Chart: Maintenance Cost and Number of Motors by Model and Type

- Displays maintenance cost per hour across various airplane models.
- The number of motors is also represented, showing how different models require different maintenance investments.

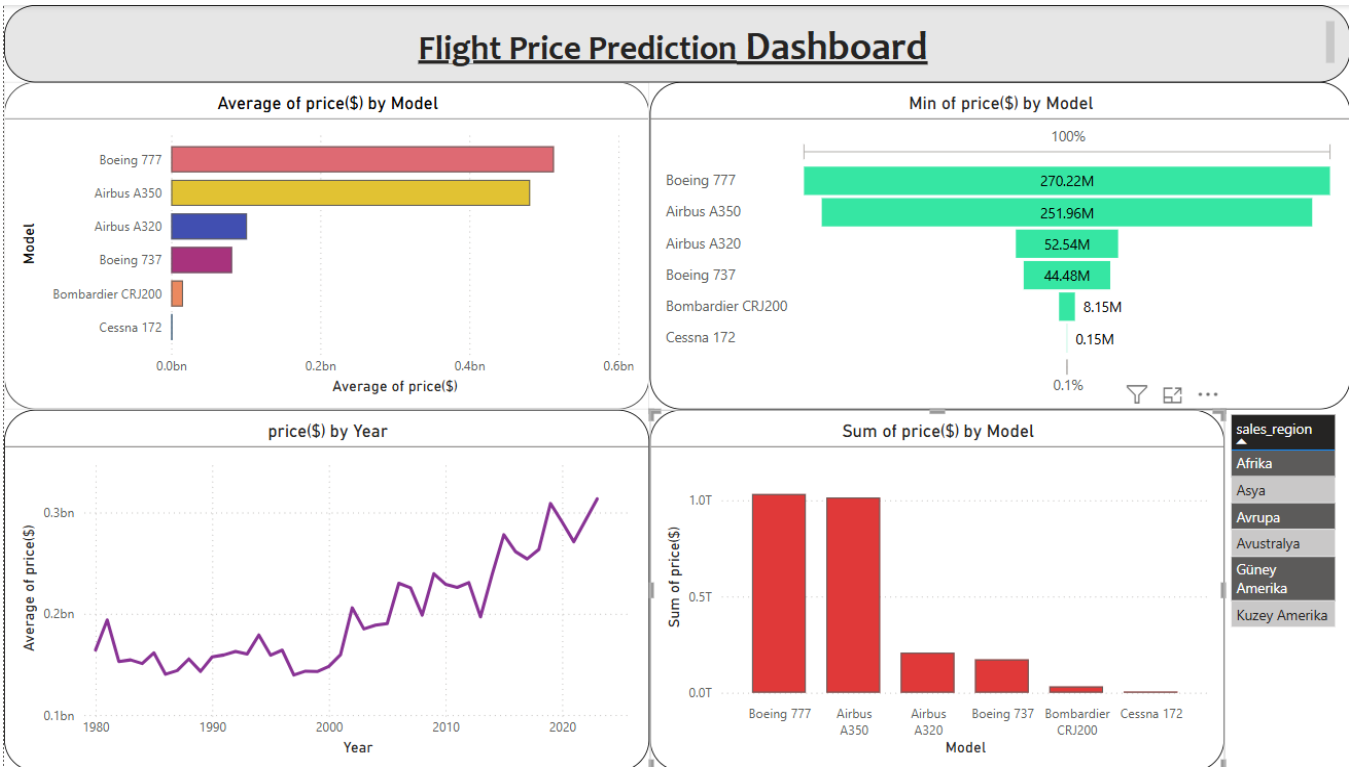
3. Pie Chart: Fuel Consumption by Model

- A pie chart showing fuel consumption distribution across different aircraft models.
- Boeing 777 has the highest fuel consumption, while smaller aircraft like Cessna 172 consume less fuel.

4. Waterfall Chart: Price Trend by Year and Model

- A waterfall chart showing price trends over time.
- It highlights the increase and decrease in prices for each model, indicating overall market trends.

Flight Price Analysis & Prediction Dashboard



This **Flight Price Prediction Dashboard** visualizes various insights about airplane prices based on different attributes such as model, year, and sales region.

Dashboard Insights:

1. Bar Chart: Average Price by Model

- Shows the **average price ($) of different aircraft models**.
- The **Boeing 777** and **Airbus A350** have the highest average prices.
- Smaller aircraft like **Cessna 172** have significantly lower average prices.

2. Bar Chart: Minimum Price by Model

- Displays the **minimum price** of each aircraft model.
- Boeing 777 has the **highest minimum price ($270.22M)**.
- Smaller aircraft like **Cessna 172** have much lower minimum prices.

3. Line Chart: Price by Year

- Shows the **trend of average aircraft prices over time**.
- The chart suggests an **increasing trend in aircraft prices** from the 1980s to 2020.
- The prices fluctuate but generally rise over time, possibly due to inflation, technological advancements, and demand changes.

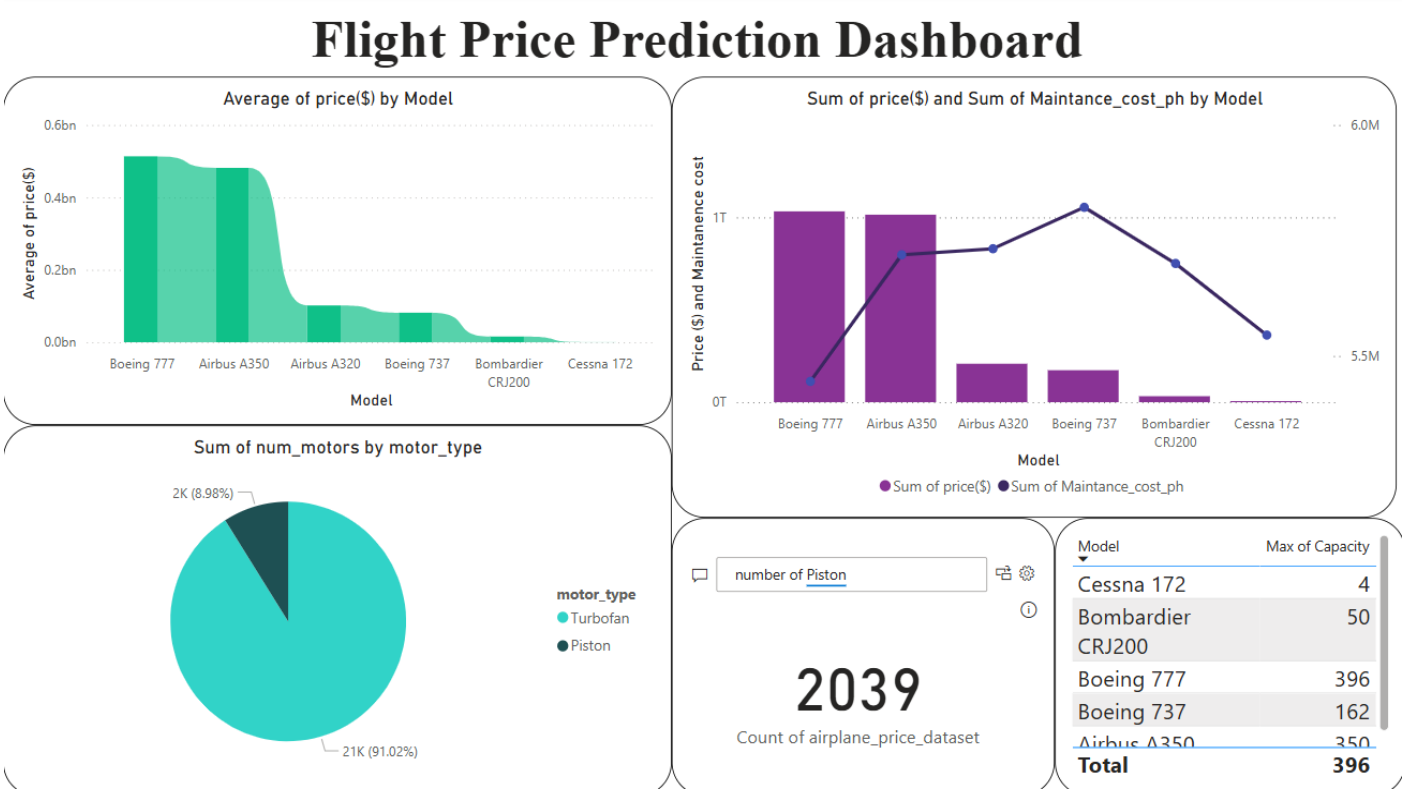
4. Bar Chart: Sum of Prices by Model

- Displays the **total sum of prices** for each aircraft model.
- **Boeing 777 and Airbus A350** dominate in terms of total price value.
- Smaller aircraft models contribute much less to the total price.

5. Slicer: Sales Region Filter

- A filter allows selecting different **sales regions (e.g., Africa, Asia, Europe, North America, etc.)**.
- Helps analyze aircraft prices based on geographical regions.

Aircraft Price & Maintenance Analysis Dashboard



This dashboard provides insights into aircraft pricing, maintenance costs, motor types, and capacity across different aircraft models. The visualizations help in understanding price trends, maintenance expenses, and engine distribution.

Dashboard Insights:

1. Bar Chart: Average Price ($) by Model

- Displays the average price of different aircraft models.
- Boeing 777 and Airbus A350 have the highest average prices, while smaller aircraft like Bombardier CRJ200 and Cessna 172 have lower prices.

2. Combo Chart (Bar & Line Chart): Sum of Price ($) & Maintenance Cost per Hour

- The bar chart represents the total price of each aircraft model.
- The line chart represents the maintenance cost per hour for different models.
- Airbus A350 and Boeing 777 have high total prices, while maintenance costs peak at mid-sized aircraft.

3. Pie Chart: Sum of Number of Motors by Motor Type

- Shows the distribution of engine types in the dataset.
- 91.02% of the aircraft use turbofan engines, while 8.98% use piston engines.
- This suggests that most aircraft in the dataset are commercial jets rather than small general aviation planes.

4. Q & A: Question and Answer

- It enables users to type questions in plain English and receive visualizations or insights based on the data.
- Ask questions about the dataset in natural language to create new visuals.
- Displays a count of 2,039 piston engine aircraft in the dataset.

5. Table: Max Passenger Capacity by Model

- Shows the maximum seating capacity for different aircraft models.
- Boeing 777 has the highest capacity (396 passengers), followed by Boeing 737 and Airbus A350.
- Smaller aircraft like Cessna 172 (4 passengers) and Bombardier CRJ200 (50 passengers) highlight the contrast in passenger capacity across different aircraft categories.

CHAPTER - V

MODEL BUILDING

Airplane Price Prediction

Predicting airplane prices is a Supervised Machine Learning problem, specifically a Regression task because the target variable (Price) is continuous. We aim to estimate Price based on input features like Seating capacity, Engine type, Number of Engines, etc. Unlike classification, where labels are discrete like "low" or "high" price, regression predicts a numeric value.

Supervised Learning

- The model learns from labeled data (input-output pairs).
- Supervised learning is the ideal approach for this project as it involves predicting airplane prices based on historical data with labeled features.
- Among various regression techniques, Multiple Linear Regression (MLR) is chosen due to its ability to model the linear relationship between multiple independent variables, such as year, number of engines, engine type, capacity, range, fuel consumption, maintenance cost, and sales region, with the dependent variable (price).
- MLR is computationally efficient, easy to interpret, and works well for structured numerical data, making it a suitable choice for this dataset.
- The model provides insights into how each factor influences airplane prices, aiding in decision-making.
- However, if the data exhibits nonlinear relationships or high multicollinearity among variables, alternative models like Decision Trees, Random Forest, or Gradient Boosting could be considered.

Regression Algorithms

There are many regression algorithms which are suitable for this problem. And the algorithms are,

- Multiple Linear Regression (MLR) → Suitable for interpreting relationships between features and price.
- Decision Trees / Random Forest → Handles nonlinear relationships and feature interactions well.
- Gradient Boosting (XGBoost, LightGBM) → Often used for high-performance predictive modeling.
- Support Vector Regression (SVR) → Effective when data has complex patterns.
- Neural Networks (Deep Learning) → Can capture complex relationships in large datasets.

5.1 Algorithm

Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a supervised learning algorithm used for predicting a continuous target variable based on multiple independent variables. It extends **simple linear regression**, which involves only one predictor, by including multiple predictors.

Equation of MLR

$Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$

Where:

- Y = Dependent variable (price of an airplane)
- X1,X2,...,Xn = Independent variables (engine type, fuel consumption)
- β0 = Intercept
- β1,β2,...,βn = Coefficients (weights for each predictor)
- ϵ (epsilon) = Error term

Predicting the Airplane Prices is a Supervised Learning problem. The **Multiple Linear Regression** Algorithm was chosen as the best fit.

5.2 Training and Test dataset

Splitting the data into a train and test set is a crucial step in building a machine learning model because it allows us to assess how well the model generalizes to unseen data. The training set is used to train the model by learning patterns from the data, while the test set is kept separate and used only for evaluation. This helps prevent overfitting, where the model learns the training data too well but fails to perform well on new data. By testing on unseen data, we can measure the model's accuracy, identify potential issues, and improve performance before deploying it in real-world applications. A typical split ratio is 80% training and 20% testing, though this can vary depending on the dataset size.

The Airplane Dataset is split into Training Set and Test set with the function called **set.seed() function** and **CreateDataPartition() function.**

```{r}
# Split data into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(data$Price, p = 0.8, list = FALSE)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]
```

- Sets a seed value for randomization to ensure reproducibility.
- This makes sure that every time the code runs, the same split occurs.
- createDataPartition() is a function from the caret package that splits data while maintaining class distribution.
- data$Price: The dataset is split based on the Price column.
- p = 0.8: Specifies 80% of the data will be used for training, and the remaining 20% for testing.
- list = FALSE: Returns an index vector instead of a list.
- Uses the indices generated to extract training data (80% of the original dataset).
- The remaining 20% of the data (not included in training) is used as test data.

5.3 Model

- A model is a mathematical representation of a real-world process that is trained on data to make predictions or decisions. It captures patterns and relationships between input (features) and output (Price).
- The model learns the best values for β using a training dataset to minimize prediction errors.
- I created a model using lm() to predict the price of an airplane using several other variables as predictors. The airplane dataset is split into two parts: 80% for the training set and 20% for the test set. First, we split the data, then train the model on the training set, and finally test it with unseen data from the test set.

- Linear regression requires one response variable and one or more predictor variables. Here, the airplane price is the response variable, while Range, Number of Engines, Engine Type, Fuel Consumption, and Age are the major predictors.

```
```{r}
# Train a linear regression model
model <- lm(Price ~ Year + Num_Engines + Engine_Type + Capacity + Range_km + Fuel_Consumption_Lph +
Maintenance_Cost_per_Hour + Sales_Region,data = trainData)
```
```

This model predicts **Price** based on multiple factors (e.g., **Year, Engine Type, Capacity, Fuel Consumption, etc.**).

Code Explanation

1. **lm()**: This function fits a **linear regression model** in R.

2. **Price ~ Year + Num_Engines + Engine_Type + Capacity + Range_km + Fuel_Consumption_Lph + Maintenance_Cost_per_Hour + Sales_Region**:

   Specifies the **dependent variable (Price)** and **independent variables (predictors)**:

   - Price: The target variable (airplane price).
   - Year: The manufacturing year of the airplane.
   - Num_Engines: Number of engines in the aircraft.
   - Engine_Type: The type of engine used (categorical variable).
   - Capacity: Passenger capacity of the airplane.
   - Range_km: Maximum flight range in kilometers.
   - Fuel_Consumption_Lph: Fuel consumption per hour.
   - Maintenance_Cost_per_Hour: The cost of maintaining the aircraft per hour.
   - Sales_Region: The region where the aircraft is sold (categorical variable).

3. **data = trainData**: Specifies that the model should be trained on the trainData dataset.

CHAPTER - VI

EVALUATION OF MODEL

6.1 Model Evaluation

Model evaluation is the process of assessing a machine learning model's performance to determine how well it makes predictions on unseen data. It helps identify issues like overfitting, underfitting, and whether the model is suitable for real-world use.

Summary of the Model

The summary() function is used after building a model in R to evaluate its performance and statistical significance. It provides key insights into the model, helping us understand how well it fits the data.

```r
# Model summary
summary(model)
```

```
Call:
lm(formula = Price ~ Year + Num_Engines + Engine_Type + Capacity +
    Range_km + Fuel_Consumption_Lph + Maintenance_Cost_per_Hour +
    Sales_Region, data = trainData)

Residuals:
       Min         1Q     Median         3Q        Max
-208842987  -38793766   -9864641   34492056  402823435

Coefficients: (1 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -7.643e+09  1.058e+08 -72.251   <2e-16 ***
Year                       3.807e+06  5.271e+04  72.229   <2e-16 ***
Num_Engines               -5.580e+07  3.498e+06 -15.952   <2e-16 ***
Engine_TypeTurbofan             NA         NA      NA       NA
Capacity                  -8.372e+05  3.731e+04 -22.440   <2e-16 ***
Range_km                   6.203e+04  9.282e+02  66.828   <2e-16 ***
Fuel_Consumption_Lph      -1.546e+04  1.165e+05  -0.133    0.894
Maintenance_Cost_per_Hour  6.526e+02  5.186e+02   1.258    0.208
Sales_RegionAsya           4.688e+05  2.308e+06   0.203    0.839
Sales_RegionAvrupa        -1.946e+06  2.300e+06  -0.846    0.398
Sales_RegionAvustralya     2.534e+04  2.286e+06   0.011    0.991
Sales_RegionGüney Amerika  1.030e+06  2.308e+06   0.446    0.656
Sales_RegionKuzey Amerika -7.745e+05  2.313e+06  -0.335    0.738
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66630000 on 9892 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.9149
F-statistic:  9675 on 11 and 9892 DF,  p-value: < 2.2e-16
```

Statistical insights

- Coefficients Table: Shows predictor significance (p-values < 0.05 indicate strong impact).
- Residual Standard Error (RSE): Measures prediction accuracy (lower is better).
- R-squared & Adjusted R-squared: Indicates how well predictors explain the target variable.
- F-statistic & p-value: Tests overall model significance.

Predict on Test data

- After training the model, you need to test how well it performs on unseen data.

- We can compare Predicted_Price with actual prices in testData to measure accuracy.

- It also Helps in calculating RMSE, R^2, MAE, etc., to assess prediction quality.

```{r}
# Predict on test data
testData$Predicted_Price <- predict(model, newdata = testData)
```

The model was evaluated on the test data using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Model Performance

Model performance refers to how well a machine learning model makes predictions on unseen data. It is evaluated using various metrics that measure the difference between predicted and actual values.

```r
# Evaluate model performance
mse <- mean((testData$Price - testData$Predicted_Price)^2)
r2 <- cor(testData$Price, testData$Predicted_Price)^2
rmse <- sqrt(mse)
mae <- mean(abs(testData$Price - testData$Predicted_Price))
accuracy <- (1 - (mae / mean(testData$Price))) * 100
```

```r
# Print evaluation metrics
cat("Mean Squared Error:", mse, "\n")
cat("Root Mean Squared Error:", rmse, "\n")
cat("Mean Absolute Error:", mae, "\n")
cat("R-squared:", r2, "\n")
cat("Accuracy Percentage:", accuracy, "%\n")
```

```
Mean Squared Error: 4.542939e+15
Root Mean Squared Error: 67401326
Mean Absolute Error: 49496987
R-squared: 0.9157956
Accuracy Percentage: 75.24992 %
```

Code Explanation

cat("Mean Squared Error:", mse, "\n")

- Prints the **Mean Squared Error (MSE)**, which measures the average squared differences between actual and predicted values.
- Lower MSE indicates better model performance.

cat("Root Mean Squared Error:", rmse, "\n")

- Prints the **Root Mean Squared Error (RMSE)**, which is the square root of MSE.
- RMSE provides an error estimate in the same unit as the target variable (e.g., airplane price).

cat("Mean Absolute Error:", mae, "\n")

- Prints the **Mean Absolute Error (MAE)**, which is the average absolute difference between actual and predicted values.
- Lower MAE indicates better accuracy.

cat("R-squared:", r2, "\n")

- Prints **R-squared (R^2)**, which measures how well the model explains variance in the data.
- Closer to **1** means a better fit.

cat("Accuracy Percentage:", accuracy, "%\n")

- Prints the estimated **Accuracy (%)**, calculated based on the MAE and mean price.
- Higher accuracy (%) is better.

Observations of the Model

1. Mean Squared Error (MSE): 4.542939e+15

- Measures the average squared differences between actual and predicted prices.
- Large values indicate significant prediction errors.

2. Root Mean Squared Error (RMSE): 67401326

- Square root of MSE, giving error in the same unit as price.
- High RMSE suggests large variations between actual and predicted prices.

3. Mean Absolute Error (MAE): 49496987

- Measures the average absolute difference between predicted and actual values.
- The model, on average, predicts prices with an error of ~49.5 million.

4. R-squared (R^2): 0.9157956

- Indicates how well the model explains price variation.
- 91.58% of the variance in airplane prices is explained by the model.
- A high R^2 suggests a good fit.

5. Accuracy Percentage: 75.25%

- Measures prediction accuracy relative to the mean price.
- A 75.25% accuracy suggests that the model performs reasonably well but still has room for improvement.

Frequency count of aircraft manufacturers

```{r}
table(data$Model)

```

The table() function is used to create frequency tables, summarizing the count of unique values in a categorical or numerical variable. Returns a table showing how many times each model appears.

```
Airbus A320      Airbus A350      Boeing 737      Boeing 777 Bombardier CRJ200      Cessna 172
       2048             2107            2121            2011             2051            2039
```

It generates a frequency count of the different airplane models in the dataset. This helps to analyze the distribution of aircraft models**.**

CHAPTER - VII

PREDICTION AND INFERENCE

7.1 Prediction

Prediction is the process of estimating unknown values based on patterns identified in historical data. In this project, the goal was to predict airplane prices using a Multiple Linear Regression model. The model was trained on a dataset containing features such as engine type, seating capacity, fuel consumption, maintenance costs, and aircraft age. By analyzing these factors, the model generates price estimates for Airplanes. This predictive approach helps Airlines, Leasing Companies, and Buyers make data-driven decisions when purchasing or selling aircraft. While the model achieved a high R-squared value (0.915), indicating strong predictive power, there is room for improvement by incorporating additional variables or using advanced machine learning techniques.

7.2 INFERENCE

Inference focuses on understanding the relationships between different variables and their influence on the target outcome. In the context of airplane price prediction, inference helps determine how factors like fuel efficiency, production year, and engine type impact pricing trends. For example, the model might reveal that Airplanes with turbofan engines generally have higher prices due to their fuel efficiency and performance advantages.

Additionally, inference helps assess the statistical significance of each variable, identifying the most influential predictors. This understanding enables industry professionals to make strategic decisions, such as optimizing aircraft specifications for better market value or forecasting future pricing trends based on economic conditions.

CHAPTER - VIII

CONCLUSION

This project focused on developing a predictive model for estimating airplane prices using Multiple Linear Regression, incorporating extensive data cleaning, preprocessing, exploratory data analysis (EDA), and data visualization techniques. The dataset, consisting of key aircraft attributes such as engine type, capacity, range, fuel consumption, and maintenance costs, was thoroughly analyzed to identify patterns and trends influencing airplane prices.

To enhance interpretability and insights, interactive dashboards were created to visualize key relationships between various aircraft characteristics and pricing. These dashboards provided a comprehensive overview of maintenance costs, fuel efficiency, engine types, and price trends across different airplane models, enabling stakeholders to make well-informed decisions.

The Multiple Linear Regression model effectively captured variations in airplane prices, identifying the most influential factors that drive aircraft valuation. Engine type, capacity, and range emerged as key determinants of price, highlighting their significance in the aviation market. The model's performance demonstrated strong predictive capabilities, offering valuable insights into how different aircraft attributes influence pricing dynamics.

To further refine the model and enhance its predictive power, future enhancements will explore non-linear models such as Decision Trees, Random Forest, and Neural Networks. Additionally, incorporating macroeconomic factors, industry trends, and real-time market data could provide even greater accuracy and robustness, making the model even more applicable to real-world scenarios.

By combining advanced machine learning techniques with data-driven insights, this project contributes to a deeper understanding of airplane pricing and helps industry professionals, buyers, and sellers make informed financial and strategic decisions.

REFERENCES

Gareth M. James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013*). An Introduction to Statistical Learning: with applications in R.*

https://www.statlearning.com/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

https://www.statlearning.com

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis.* Wiley.

https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis%2C+6th+Edition-p-9781119578722

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer.

https://hastie.su.domains/ElemStatLearn/